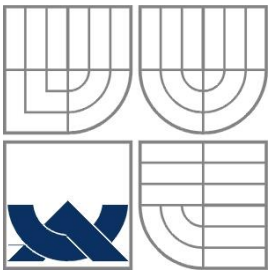


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

URČOVÁNÍ AUTORSTVÍ

AUTHORSHIP IDENTIFICATION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

ONDŘEJ FABIÁNEK

VEDOUCÍ PRÁCE
SUPERVISOR

Doc. RNDr. PAVEL SMRŽ, Ph.D.

BRNO 2016

Abstrakt

Tato práce se zabývá určováním autorství na základě znalosti předchozích textů autora. Cílem bylo prostudovat existující metody pro určování autorství a následně vytvořit systém, který dovede s vysokou pravděpodobností identifikovat autora textu. Zaměřuji se zde na určování autorství anglicky psaných knih a součástí řešení je též grafické rozhraní.

Abstract

This bachelor's thesis deals with authorship identification based on knowledge of author's previous texts. The aim is to analyze existing methods of authorship attribution and create a system, which is capable of highly successful authorship identification. The system is based on a multivariate analysis and specializes at English books. Part of the solution is also a graphic user interface.

Klíčová slova

autorství, knihy, stylometrická analýza, zpracování přirozeného jazyka, anonymní dokument

Keywords

authorship, books, stylometric analysis, natural language processing, anonymous document

Citace

FABIÁNEK, Ondřej: *Určování autorství*. Brno, 2016. 31s. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Smrž Pavel.

Určování autorství

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vytvořil samostatně pod vedením pana doc. RNDr. Pavla Smrže, Ph.D.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……
Ondřej Fabiánek
13. května 2016

Poděkování

Děkuji tímto vedoucímu mé práce za všechny poskytnuté cenné rady.

Obsah

Kapitola 1

Uvod

Tato práce se zabývá zkoumáním metod používaných k určování autorství textů. Jejím cílem bylo obeznámit se s existujícími postupy pro identifikaci autora pomocí statistického rozboru jeho textů, experimentovat s použitím různých postupů a následně s využitím získaných poznatů vytvořit systém, který by s vysokou mírou úspěšnosti dokázal identifikovat autora na základě znalosti jeho předchozích textů. Různé typy textů (emaily, chaty, …) vyžadují pro dosažení maximální efektivity odlišné strategie. Tato práce se soustřeďuje na určování autorství u anglicky psaných knih. Cíleným uplatněním tohoto je pak identifikace autorů, píšících pod různými pseudonymy.

Metody určování autorství vycházejí z hledání podvědomých vzorů, kterých se každý autor při psaní dopouští. Někteří autoři často používají nějaké atypické slovo, jiní tíhnou k psaní dlouhých vět, žádná takováto charakteristika však není zcela spolehlivá a obvykle se proto používají kombinace většího množství z nich. Prvním ze dvou hlavních problémů, kterými se tedy tato práce zaobírá, je hledání nejvhodnější skupiny těchto charakteristik. V kapitole 2 se lze dočíst o historickém vývoji určování autorství. Jaké charakteristiky byly používány před vynalezením počítače a jaký vliv mělo jeho vynalezení na vývoj této vědní disciplíny. Čtenář je zde obeznámen se současným stavem této problematiky a jejích možných uplatnění. Kapitola 3 obsahuje podrobný popis některých vybraných příznaků.

Druhým problémem, kterému se tato práce velmi podrobně věnuje, je způsob porovnávání naměřených charakteristik. V kapitole 4 jsou diskutovány možnosti použití metod strojového učení či jiných postupů, využitelných k tomuto účelu.

Kapitola 5 se zabývá návrhem programu. Věnuje se praktickým aspektům použití tohoto programu a zmiňuje externí nástroje, vhodné k usnadnění řešení některých problémů. V kapitole 6 lze nalézt popis grafického rozhraní a jeho výslednou podobu. Kapitola 7 pak popisuje proces shromáždění testovacích dat a způsob testování přesnosti programu. Předposlední a nejobsáhlejší je kapitola 8. Zabývá se vyhodnocením výsledků, dopodrobna rozebírá jednotlivé použité charakteristiky, metody a informuje o jejich vlivu na výslednou přesnost. V závěru této práce jsou zmíněny některé možné úpravy, jejichž realizací by mohlo dojít k dalšímu zpřesnění výsledků.

Kapitola 2

Určování autorství

V této kapitole je prezentován souhrnný pohled na celou problematiku určování autorství. Bližší pozornost je zde věnována také motivaci pro identifikaci spisovatelů, píšících pod pseudonymy.

2.1 Historický vývoj

Řešení sporů o autorství textů je problém, který lidé řešili již dlouho před vynálezem počítače. První zdokumentované snahy pochází z roku 1787 [1]. Edmond Malone, učenec a odborník na Shakespereova díla, tehdy zpochybnil autorství tří částí Shakespearovy tetralogie Henry IV. Odůvodňoval to řadou atypických vlastností rýmu, nevyskytujících se v ostatních dílech tohoto autora. V návaznosti na tuto studii se pak další učenci začali věnovat zkoumání autorství Shakespearových děl, zabývali se však téměř výhradně vlastnostmi rýmu.

V roce 1887 univerzitní profesor T. C. Mendenhal poprvé realizoval metodu, navrženou o 26 let dříve Augustus de Morganem [12]. Jeho postup pro identifikaci autora byl založen na myšlence, že pokud nakreslíme histogram délek slov pro dvě knihy, budou si velmi podobné, jestliže jsou napsány stejnou osobou, a odlišné, pokud ne. Výsledky jeho experimentů však ukázaly, že rozložení délek slov bylo mezi různými autory většinou velmi podobné a jako prostředek pro správné určení autora byly použitelné jen u několika osob. Kvůli nutnosti manuálně počítat délky všech slov bylo v této době nemožné ověřit důvěryhodnost této metody na větším vzorku dat, nicméně T. C. Mendenhal touto svou prací položil základy moderní stylometrii, neboli statistickému přístupu k určování autorství.

V první polovině 20. století se dále zkoumaly různé variace délek slov, jakožto i řada nových charakteristik, jako například délka vět, frekvence písmen či výskyt specifických slov. Největší rozmach však tato vědní disciplína zažila po vynalezení počítače, který umožnil testování nejrůznějších hypotéz ve velmi krátké době a na velkém množství dat. V roce 1976 byl posudek, založený na stylometrii, použit jako důkaz u soudu. Reverend Andrew Morton svědčil jakožto expert na stylometrii ve prospěch obhajoby, že sepsané přiznání bylo dílem několika autorů a tím pádem zfalšované policií. V 90. letech začaly soudy v Americe výrazně omezovat použití posudků, založených na stylometrii v důsledku vlny kritiky spolehlivosti těchto metod ze strany vědecké obce. V dnešní době je posudek, založený na stylometrii v některých státech jako důkaz u soudu přípustný, nicméně soudy se k jeho použití uchylují jen velmi výjmečně a je to podmíněno tím, že bude pro všechny užité metody experimentálně ověřeno a v posudku uvedeno, jaká je pravděpodobnost chybného závěru [2].

2.2 Řešené problémy

Přesnost výsledku, dosaženého programem na určování autorství, je dána strategií přístupu k řešení 2 základních problémů. Těmi jsou volba vhodných příznaků a metoda porovnávání příznakových

vektorů. V případě volby příznaků je třeba zohlednit typ textů, se kterými pracujeme. Analýza některých příznaků může vyžadovat určité dodatečné nástroje, jako jsou například slovníky, značkovače slovních druhů či jiných významových aspektů textu. Důvodem nejednoduchosti řešení úloh určování autorství je velký vliv délky a tématu textu na většinu příznaků [3]. Ačkoli jsou některé označovány jako nezávislé na tématu, tak ani tyto od něj nejsou odproštěny docela.

Protože je určování autorství postaveno na identifikaci vzorů, jichž se člověk dopouští podvědomě, není těžké záměrně napsat takový text, který bude i těmi nejúspěšnějšími programy přisouzen někomu jinému.

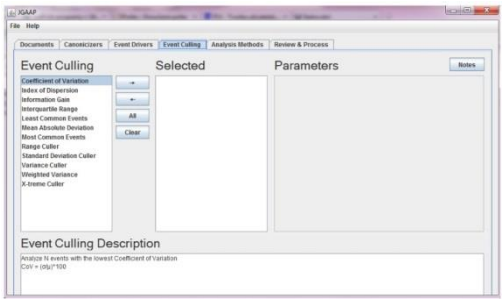
2.3 Spisovatelé píšící pod pseudonymem

Existuje řada důvodů, proč se někteří spisovatelé rozhodnou skrývat pod více různými jmény. Stephen King publikoval několik knih pod jménem Richard Bachman, aby nebyl trh „přesycen“ jeho knihami a protože si chtěl údajně ověřit, jestli za svůj úspěch skutečně vděčí svým dovednostem, nebo jen štěstí. Po vydání 5 knih pod tímto pseudonymem si jistý čtenář všiml podobnosti mezi stylem Kinga a Bachmana, načež vyhledal v kongresové knihovně dokumenty, potvrzující, že se v obou případech jedná o knihy S. Kinga [4].

Dalším známým případem odhalení totožnosti autora, píšícího pod jiným jménem, je kniha Volání kukačky. Jako její autor byl uváděn Robert Galbraith, ačkoli byla napsána J. K. Rowlingovou. Pseudonym použila proto, aby se nemusela potýkat s velmi vysokými nároky na jakoukoli její novou knihu. Poté, co skrze jistého právníka pronikla do tisku informace o tom, že tuto knihu napsala ona, bylo zapotřebí tuto informaci ověřit. K tomuto účelu byl použit program pro určování autorství JGAAP[1], který potvrdil její pravdivost [5].

2.4 Existující systémy pro určování autorství

Ačkoli je určování autorství značně diskutované téma a bylo již vytořeno velmi mnoho programů či skriptů pro tyto účely, většina jich je velmi těžko dohledatelná. Výjimku tvoří 2 systémy, které naprosto dominují výsledkům všech vyhledávání. Jedná se o výše zmíněný program JGAAP a webové rozhraní *aicbt*[2].



Obrázek 1: Volně dostupný program pro určování autorství JPAAG v6.0

[1] http://*evllabs.com/jgaap/w*
[2] http://www.aicbt.com/authorship-attribution/online-software/

Kapitola 3

Příznaky relevantní pro určování autorství

Tato kapitola se zabývá výčtem všech podstatných příznaků, které lze analyzovat za účelem definování autorova stylu. Z nastudované literatury vyplynulo, že přínosnost jednotlivých příznaků se může napříč autory velmi lišit. Zatímco některý autor může ve svých textech stabilně udržovat například stejnou průmernou délku vět, tak u jiného může docházet k jejím významným výkyvům [6]. Lze však říci, že u některých příznaků bývají tyto výkyvy větší a vyskytují se častěji, než u jiných. V minulosti byly postupně vymýšleny, testovány a zatracovány nejrůznější příznaky, které samy o sobě nebyly dostatečně přesné na to, aby se podle nich dal autor spolehlivě určit. V dnešní době se již tento přístup nepoužívá a převládá tzv. Multivariate analysis (neboli přístup založený na celé řadě příznaků) [12].

Příznaky jsou zde členěny do 3 kategorií. Vzhledem k tomu, že již byly vyzkoušeny stovky, ne-li tisíce příznaků či jejich variant, není možné věnovat zde prostor všem a následuje proto výčet pouze vybrané skupiny z nich.

3.1 Obsahově závislé příznaky

Ačkoli má téma textu určitý vliv na většinu, ne-li všechny charakteristiky, které kdy byly navrženy, některé jsou jím ovlivňovány výrazněji, než jiné. Tyto příznaky je pak vhodné používat pouze k porovnávání takových textů, o kterých víme, že jsou si tématicky velmi blízké. Nehrozí pak, že by převážila informace o obsahu textu nad informací o stylu autora. Následuje popis několika takových.

3.1.1 Velikost slovní zásoby

Pro analyzování tohoto příznaku je zapotřebí mít k dispozici nástroj, který dokáže slova normalizovat (odstraňovat u slov koncovky -ing, -ed, -s, -es, apod.), abychom nepočítali stejná slova v různých tvarech vícekrát. Jedno z možných pojetí tohoto příznaku je to, že se na určitém úseku textu spočítají všechna slova a poté se podělí počtem různých slov. Navrhovaných postupů však existuje více. V případě, že pracujeme s příliš krátkým textem a nemáme tedy dostatečné množství slov k použití předešlé metody, jedněch z nejlepších výsledků pak lze dosáhnout výpočtem pomocí následujícího vzorce [1]:

$$LN = (1 - V^2) / (V^2 \log(N))$$

N zde udává celkový počet slov na určitém úseku textu a *V* množství různých slov na stejném úseku.

3.1.2 Relativní překrytí slovní zásoby

Tato metoda měří, do jaké míry se překrývají slova užitá ve dvou různých textech a na tomto základě se snaží určit, zda jsou od téhož autora. Pro dva různé autory může být pomocí některé z variant výše,

naměřena podobná relativní velikost slovní zásoby, avšak oba mohli při psaní aktivně používat jiné části své slovní zásoby, což může být schopna odhalit tato metoda.

3.2 Obsahově nezávislé příznaky

Do této kategorie spadají ty příznaky, které téma obsahu daného textu ovlivňuje pouze v malé míře. Následuje jejich výčet.

3.2.1 Funkční slova

Jedná se o skupinu slov, která jsou nezávislá na obsahu. Spadají sem především různé spojky, zájmena a částice. Relativní četnost jednotlivých funkčních slov nám pak udává informaci o autorově stylu. Je to jeden z nejspolehlivějších příznaků [3]. Práce, které v minulosti využívaly tohoto příznaku k určování autorství, však nejsou jednotné ohledně volby konkrétního počtu a volbě funkčních slov. Důležitým faktorem je také způsob, jakým se relativní četnost vypočítá. Jedním z možných přístupů je tento:

$$C_i = \frac{v_i}{W}$$

Kde C_i představuje relativní četnost jednoho funkčního slova, v_i celkové množství výskytů tohoto slova a W celkový počet výskytů všech funkčních slov. Jiný přístup zas jako W dosazuje celkový počet všech slov.

Zajímavá je také varianta, navržená Alvarem Ellegarem (1962). Tato metoda dělí slova na tzv. plusová a mínusová. Plusová jsou ta, jejichž poměr vůči délce textu je vyšší, než je průměrná hodnota tohoto poměru vysledovaná v textech velkého množství jiných autorů. Mínusová pak ta slova, která mají tento poměr nižší. Takto lze nalézt slova, která autor používá oproti jiným lidem nadprůměrně často, a slova, která v porovnání s ostatními lidmi používá naopak neobyčejně zřídka.

Za zmínku stojí i práce z r. 1993, kdy byly použity trochu neobvyklé poměry slov: *did / (did + do), no / (but + by + for + no + not + so + that + the + to + with), no / (no + not), 'to the' / to,* a *upon / (on + upon).*

3.2.2 Délka vět

Měřit lze buď průměrnou délku vět, či pracovat s histogramem rozložení jejich délek. Jednotkou pro měření délky vět mohou být buď slova, nebo znaky. Můžeme ji měřit s omezením (např. pouze relativní množství šestislovných a sedmislovných vět), či bez omezení. Z literatury vyplynulo, že nejlepších výsledků bývá dosaženo při použití bez omezení a se slovem, jakožto jednotkou. Tato metoda je vhodná pouze v kombinaci s jinými, neboť neumožňuje dostatečně velkou diferenciaci u většího počtu autorů. Její užití bývá obvykle vhodné v kombinaci s funkčními slovy.

3.2.3 Délka slov

Zde lze opět pracovat buď s jedním údajem, udávajícím průměrnou délku slov, nebo s histogramem, kde máme údaj o počtu slov o délce jednoho znaku, dvou znaků, etc. Délku slov lze měřit v počtu

písmen či v počtu samohlásek. Stejně jako v případě vět je zde nejvýhodnější měřit spektrum všech délek slov a neomezovat se pouze na některé. Studie prováděné v minulosti nejsou jednotné v tom, zda-li je použití tohoto příznaku žádoucí. Z většiny zdrojů vyplynulo, že není, avšak existuje též studie (z roku 2002), ve které délka slov dosáhla lepších výsledků, než délka vět [1].

3.2.4 N-tice znaků

Výhodou přístupu postaveného na tomto příznaku je to, že je nezávislý na jazyce a že zohledňuje i používání čárek ve větě. Princip spočívá v tom, že na obsah souboru nenahlížíme jako na posloupnost slov, ale jako na posloupnost znaků. Pokud bychom za 'n' zvolili 3, bylo by spojení "After you" rozděleno na: |Aft|, |fte|, |ter|, |er_|, |r_y|, |_yo|, |you|. Nevýhodou tohoto přístupu je to, že je vysoce paměťově i časově náročný. Je totiž zapotřebí pamatovat si počty výskytů všech těchto n-tic a pak tyto rozsáhlé množiny porovnávat s podobnými množinami ostatních autorů, což je při velkém množství textů o délce knih velmi zdlouhavé. Tento problem lze zmírnit použitím pouze na určitém výřezu zkoumaného textu.

3.2.5 Frekvence písmen

Ačkoli byla tato charakteristika velmi dlouho ignorována, experimenty provedené T. Merriamem (1988) ukázaly, že frekvence výskytu pro různá písmena je napříč autory velmi podobná, avšak napříč texty stejného autora bývá tato frekvence ještě podobnější. Porovnáním relativní četnosti písmene *O* u Shakespeara a v hrách Christophera Marlowa Merriam zjistil, že relativní četnost *O* u všech 36 her Shakespeara byla vyšší, než 0.078, zatímco 6 ze 7 her Marlowa ji mělo nižší, než 0.078. V experimentu z roku 2005, kde byly použity různě dlouhé texty (nejkratší měl 119 slov), byl pomocí této charakteristiky v její nejjednodušší podobě (relativní výskyt všech písmen v abecedě) správně určen autor textu ve 25 % případů. Vybíralo se ze 40 možných autorů a v úspěšnosti tato charakteristika předčila délku slov, délku vět i velikost slovní zásoby.

Důvodem, proč byl tento příznak velmi často opomíjen je fakt, že nikdo nebyl schopen přesně objasnit, jak je možné, že funguje. Co způsobuje, že se vynořují vzory společné pro většinu textů jednoho autora, avšak odlišné od textů autora jiného. Spekulovalo se o řadě důvodů. Například fonetická preference autora, či v případě Shakespeara jeho preference pro jména obsahující O (Othello, Romeo, Antony, Cleopatra, etc..), avšak dodnes toto není zcela objasněno[1].

I pro tento příznak byla postupem času navržena řada modifikací. Zkoušelo se měřit pouze písmena, kterými slova začínala, nacházela se pouze na n-té pozici ve slově apod. Výpočet probíhal tak, že se pro každé písmeno vydělil počet všech slov, obsahujících toto písmeno na dané pozici, celkovým počtem všech slov. Většina těchto úprav nicméně vedla ke zhoršení výsledků. Významný nárůst přesnosti však způsobilo sledování více pozic ve slově najednou. Výše zmíněný experiment s úspěšností 25 % pro všeobecnou relativní četnost písmen dosáhl úspěšnosti 49 % při sledování prvních a posledních 6 pozic ve slově. Postup byl stejný, jako když měříme frekvenci písmen na n-té pozici (popsané výše), avšak zopakujeme to pro každou ze 12 pozic ve slově. Pro písmeno *A* bychom tedy získali jeho frekvenci na pozici 1, 2, etc..

3.2.6 Poměry slovních druhů

Tento přístup vyžaduje slovník slovních druhů či nějaký nástroj, pro jejich určování. Vychází z myšlenky, že každý autor má tendenci tíhnout k určitému pro něj typickému rozložení slovních druhů. Řada studií důvěryhodnost tohoto příznaku napadla a označila jej za nespolehlivý, nicméně určitý vzor napříč texty stejných autorů nalezen byl. Jistý přínos by tedy v tomto příznaku být mohl, avšak pouze v kombinaci s jinými.

3.2.7 N-tice slovních druhů

Zde je opět zapotřebí nějaký nástroj pro identifikaci slovních druhů. Nejobvyklejší je použití dvojic či trojic slovních druhů. Možnou modifikací je sledování pouze slovních druhů nacházejících se bezprostředně před a za nějakým zvoleným slovním druhem.

3.2.8 Interpunkce

Měřit můžeme četnost výskytu vybrané skupiny interpunkčních znamének. Výpočet probíhá vydělením množství daného typu znaménka celkovým počtem slov, znaků či všech interpunkčních znamének. Ačkoli byl tento příznak dlouhou dobu přehlížen, výsledky některých studií naznačují, že se jedná o použitelný ukazatel autorova stylu.

3.2.9 Pozice slov

Zkoumat lze to, jaká slova se obvykle vyskytují na první, poslední či všeobecně n-té pozici ve větě. Relativní výskyt určitého slova na dané pozici získáme vydělením počtu vět s tímto slovem v této pozici celkovým počtem všech vět. Některé studie používaly i kombinaci několika pozic. Četnost se pak počítá pro každou pozici zvlášť. Sledování slov na první pozici ve větě obvykle vede k mnohem lepším výsledkům, než sledování slov na konci věty. Nejlepším výsledkům z těchto variant, se kterými jsem se v literatuře setkal, bylo dosaženo pomocí sledování 4 prvních slov v každé větě. Úspěšnost takovéhoto postupu mírně předčila výsledky, jakých bylo dosaženo měřením délky vět.

3.2.10 Používání synonym

Tento příznak spočívá v měření toho, do jaké míry autor používá v textu synonyma, či jaká synonyma preferuje. Pro první variantu je zapotřebí slovník synonym. Druhá vyžaduje pouze vytvoření několika skupin synonym (*should* x *ought to; giant* x *huge* x *enormous* x *colossal;…*) a měření toho, která z těchto slov autor preferuje.

3.3 Příznaky závislé na formě

Sem lze zařadit takové příznaky, které lze zkoumat pouze u určité úzké skupiny textů. Například u emailové komunikace můžeme zkoumat, jaký je použitý pozdrav a rozloučení, jaká je délka jednotlivých odstavců apod. V případě prózy by sem mohla být zařazena kupříkladu existence prologu na začátku knihy či délka poslední kapitoly. Spolehlivost takovýchto příznaků se liší případ od případu.

Kapitola 4

Metody porovnávání příznakových sad

Neméně významnou, než volba vhodných příznaků, bývá při určování autorství též volba vhodné metody pro hledání podobností mezi příznaky nalezenými u dvou autorů. Ačkoli většina nejmodernější programů k tomuto účelu využívá strojového učení a o Support Machine Learning bylo prokázáno, že je pro tyto účely alespoň tak úspěšné, jako jakýkoli jiný algoritmus [10], velmi dobrých výsledků lze dosáhnout též pomocí metod, které strojového učení nevyužívají. V této kapitole bude zmíněn princip fungování nejběžnějších metod spolu s jejich výhodami a nevýhodami.

4.1 Různá paradigmata použití

Kromě nejběžnější varianty, kdy máme anonymní text a hledáme k němu autora, existují i odlišná paradigmata, se kterými je možné se setkat. Můžeme mít kupříkladu množinu, obsahující větší množství anonymních textů, a nemít k dispozici žádnou množinu potencionálních autorů. V takovéto situaci se lze pokusit o roztřídění těchto textů tak, aby se k sobě shlukovaly ty, které byly napsány stejnou osobou. Další situací, se kterou se lze setkat, je rozhodování o tom, kolik autorů se podílelo na napsání nějakého textu.

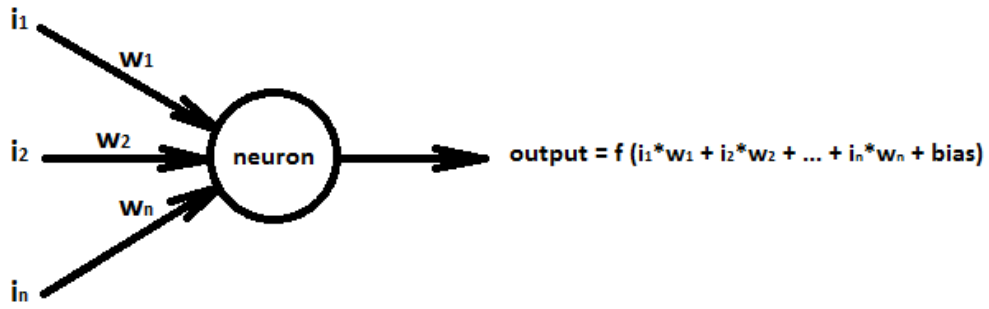
Z hlediska způsobu extrakce množiny příznaků lze nalézt 3 základní přístupy. Při určování autorství totiž vždy musíme mít k dispozici nějaký anonymní text, dále pak množinu potenciálních autorů a množinu textů, jejíž autorství je nám známo a lze jej přiřadit někomu z množiny autorů. Přístupy se liší v tom, zda-li každou knihu autora rozdělíme na kapitoly a budeme tedy mít pro ni tolik vektorů příznaků, kolik obsahuje kapitol, nebo zda budeme pracovat s knihou jako celkem a mít pro každou pouze jeden vektor. Třetí variantou je pak hybridní přístup, který je kombinací obou zmíněných. Nevýhodou udržování více sad příznaků pro každou knihu je vyšší paměťová i časová náročnost (obzvláště znatelná pokud vybíráme autora mezi stovkami osob) a fakt, že některé příznaky potřebují delší text na to, aby byly schopné konvergovat k hodnotám, typickým pro tohoto autora.

4.2 Klasický přístup

Klasickým přístupem se zde rozumí porovnávání vektorů příznaků bez použití algoritmu strojového učení. Nejjednodušší přístup pak bere složky vektoru anonymního textu (a stejně tak ostatních textů) jako souřadnice bodu v prostoru, pro všechny známé texty vypočte jejich vzdálenost od tohoto bodu a za nejpravděpodobnějšího autora nějakého textu označí toho, jehož kniha se nachází nejblíže. Sofistikovanější metody pak mohou pro různé složky používat jiný způsob výpočtu, různým složkám přikládat různé váhy apod. Výhodou tohoto přístupu oproti strojovému učení je to, že lze snadno zjistit, jaký příznak měl jaký vliv na výsledek. Které aspekty stylu anonymního textu nějakému autorovi odpovídaly a které naopak byly v jeho neprospěch. Všeobecně poskytuje širší možnosti parametrizace a větší kontrolu nad celým procesem identifikace autora.

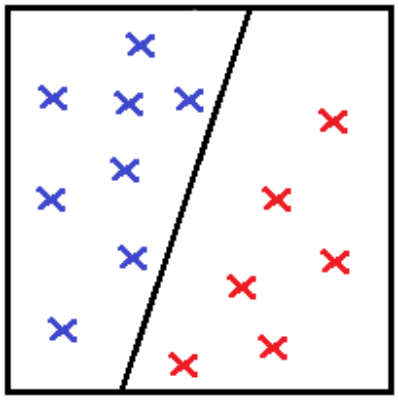
4.3 Neuronové sítě

Neuronová síť v informatice představuje soubor propojených jednotek, nazývaných „neurony“. Ty pak pomocí matematických funkcí a kolekce vah simulují fungování neuronů v mozku. Nejjednoduším modelem neuronové sítě je tzv. „perceptron“. Sestává z 1 neuronu a kolekce vážených vstupů. Výstup tohoto neuronu je funkcí sumy (tzv. „aktivační funkce) všech vážených vstupů. Nejčastěji používanou aktivační funkcí je sigmoid: $f(x) = \frac{1}{1+e^{-x}}$. [7].

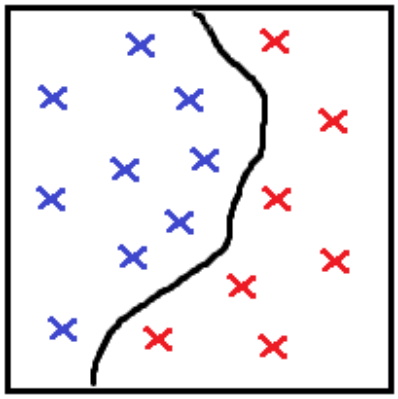


Obrázek 2: Perceptron

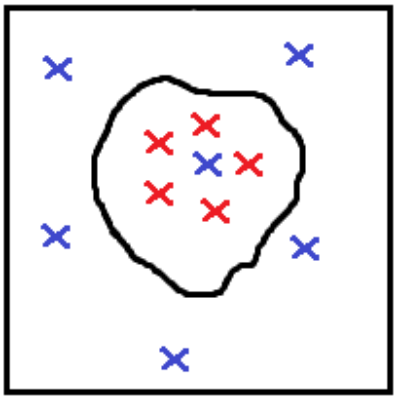
K řešení většiny problémů jsou zapotřebí sítě sestávající z většího množství neuronů, obvykle rozdělených do několika vrstev. Vrstvy umístěné mezi první a poslední se nazývají „skryté“. Zapojení může být jednosměrné, kdy výstup každého neuronu je přiveden na vstup všech neuronů následující vrstvě, či obousměrné (např. Hopfieldova síť). V případech, kdy je aproximovaná funkce spojitá, si vystačíme s jedinou skrytou vrstvou. V ostatních případech jich pak může být zapotřebí 2 či více.



Obrázek 3: Situace řešitelná bez skryté vrstvy



Obrázek 4: Situace řešitelná 1 skrytou vrstvou



Obrázek 5: Situace vyžadující 2 skryté vrstvy

Obrázky č. 3 až 5 demonstrují klasifikační úlohy řešitelné neuronovýma sítěma s různým počtem skrytých vrstev. Červené a modré křížky mohou reprezentovat texty dvou různých autorů.

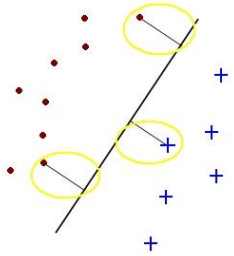
Před jejím použitím je třeba neuronovou síť natrénovat. Existuje učení s dozorem a bez dozoru. V případě učení s dozorem jsou síti poskytnuty vstupy (např. složky vektoru charakteristik, naměřených u nějakého textu) a požadované výstupy (čísla, se kterými asociujeme jednotlivé autory).

Následně jsou poupraveny váhy tak, aby se minimalizoval rozdíl mezi chtěným výsledkem a skutečným výsledkem. V případě učení bez dozoru jsou síti poskytnuty pouze vstupy a váhy jsou pak měněny tak, aby podobné vstupy vedly k podobným výsledkům. Jedna iterace poskytnutí vstupů síti a úpravě jejích vah se nazývá „epocha". Obvykle je zapotřebí mnoho epoch k natrénování sítě.

Nevýhodou neuronových sítí je jejich náchylnost k přeučení (overfitting). Při příliš malém počtu neuronů se síť neučí dostatečně detailně, při příliš velkém počtu se již naopak učí až příliš bezvýznamné detaily. S narůstajícím množství neuronů a vrstev také významně narůstá výpočetní náročnost. Není tedy praktické aplikovat neuronové sítě na vektor naměřených charakteristik, mající stovky či tisíce složek.

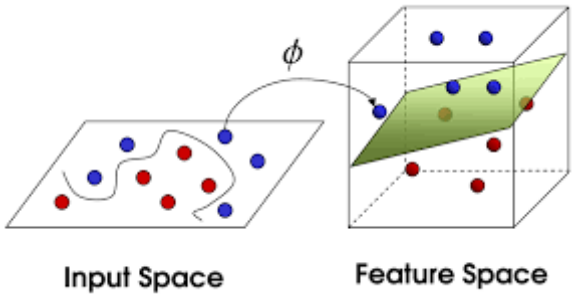
4.4 Support Vector Machines

Výhodou Support Vector Machines (SVM) oproti neuronovým sítím, je jejich schopnost vypořádat se velice rychle i s velmi velkým množstvím charakteristik [10]. Vektor příznaků může obsahovat tisíce složek a stále to nebude představovat problém. Tato metoda optimálním způsobem rozděluje prostor na dvě části. Na obrázku č. 6 lze vidět nadrovinu, dělící dvoudimenzionální prostor.



Obrázek 6: Optimální umístění nadroviny ve 2D prostoru[3].

Křížky a tečky mohou představovat texty 2 různých autorů. SVM nadrovinu umístí vždy tak, aby se maximalizovala její vzdálenost od nejbližších textů obou autorů. Liší se tak od neuronových sítí, které nemusí nalézt optimální polohu (viz obrázek č. 3). V případě, kdy data nejsou lineárně oddělitelná, SVM umožňuje provést transformaci jádra. Tato operace mapuje body do prostoru vyšší dimenze, což je učiní lineárně oddělitelnými (obrázek č. 7).



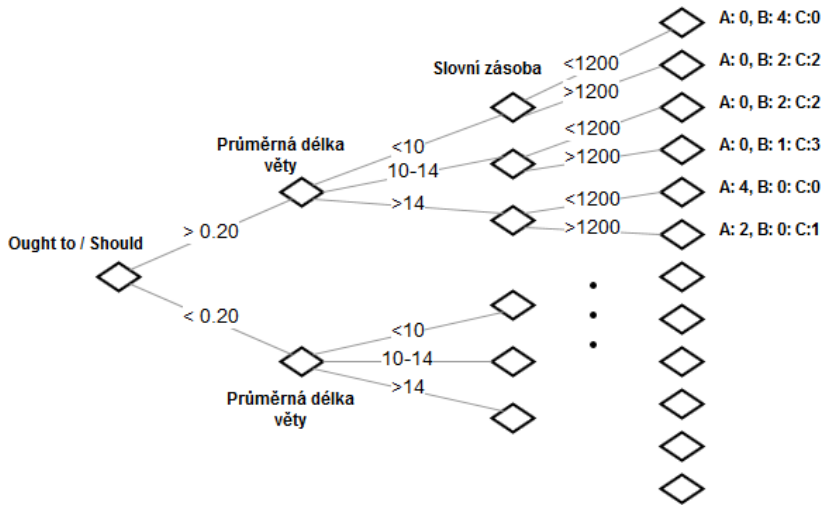
Obrázek 7: Transformace jádra[4]

[3] https://www.projectrhea.org/rhea/images/4/40/Lec11_sv_pic2_OldKiwi.jpg
[4] http://i.stack.imgur.com/1gvce.png

Pro maximalizaci přesnosti klasifikace je zapotřebí zvolit vhodný typ jádra, jeho parametry a penalizační parametr *C*. Vstupní data je vhodné normalizovat např. přemapováním hodnot do intervalu *<0;1>*, aby hodnoty nějaké složky nepřevážily hodnoty všech ostatních.

4.5 Náhodné lesy

Náhodné lesy (Random forests) jsou další klasifikační technika, kterou v minulosti úspěšně aplikovaly některé studie pro určování autorství. Sestává z kolekce rozhodovacích stromů, ty jsou zkonstruovány s použitím náhodné podmnožiny trénovacích dat (obvykle o velikosti 50-65 % celé trénovací množiny). Ukázku malého rozhodovacího stromu lze vidět na obrázku č. 8. Pokud chceme nějaký vstup přiřadit některé ze tříd A, B, nebo C, tak postupujeme grafem od kořene k listům. Pro každý list známe počet trénovacích vzorků od každé třídy, které v něm skončily. Pokud tedy s neznámým vstupem dojdeme k listu s hodnotami A = 0, B = 1 a C = 3, tak se 75% jistotou jedná o prvek třídy C a se 25% jistotou o prvek třídy B.



Obrázek 8: Rozhodovací strom

Výhodou rozhodovacích stromů je jejich schopnost identifikovat nejvýznamnější složky vstupních vektorů (tyto jsou pak umístěni nejblíže ke kořeni) a jejich schopnost pracovat jak se spojitými, tak i s diskrétními daty. Nevýhodou rozhodovacích stromů je nízká schopnost zevšeobecňovat. S rostoucí velikostí stromu narůstá riziko přeučení a snižuje se tak pravděpodobnost správného určení neznámého vstupu. Tento problém do značné míry zmírňují náhodné lesy, které průměrují výsledky velkého množství stromů a dosahují tedy lepších výsledků.

Kapitola 5

Návrh programu

Tato kapitola se z hlediska návrhu zabývá správou knih a uvádí výčet charakteristik, vybraných pro implementaci, včetně zvoleného postupu pro jejich porovnávání. Správou knih je zde míněn způsob přidávání či odebírání knih z trénovací sady a zacházení s naměřenými údajemi.

5.1 Správa knih

Ke každé knize z trénovací sady je potřeba moci přiřadit k ní jejího autora. Toho lze docílit například tak, že by se zavedl nějaký povinný formát názvů knih, který by umožňoval vyčíst jméno autora z názvu knihy. Uživatel by pak musel všechny soubory před jejich umístěním do učící sady přejmenovávat. Další možností by mohlo být přidávání knih do učící sady skrze grafické rozhraní programu. Tento způsob by však vyžadoval, aby uživatel při přidávání každé knihy uvedl jméno jejího autora, což se u větších počtů textů jeví jako nepraktické. Nakonec jsem jako nejvhodnější vybral manuální roztřídění textů do složek s názvy jejich autora. Adresář s korpusem bude tedy obsahovat pro každého autora 1 složku a v ní libovolně uspořádané a pojmenované texty daného autora. Jediný požadavek je ten, že texty musí mít koncovku .txt.

Druhým problémem k vyřešení je zde to, že pokaždé, když je třeba nalézt autora nějakého anonymního textu, je zapotřebí informací získaných z analýzy všech textů. V případě dlouhých textů, jako jsou knihy, by bylo testování neuskutečnitelné, pokud bychom měli vždy znova procházet všechny knihy. Z tohoto důvodu bylo rozhodnuto všechny informace, získané z analýzy knih, ukládat na disk. V každé složce s autorem se po prvním zpracování knih vytvoří soubor, ve kterém budou všechny relevantní informace extraktované z jeho knih. Každá kniha je tedy analyzována pouze jedenkrát a následně již program pracuje pouze s tímto souborem. Použití programu pak bude zahrnovat jedno velmi zdlouhavé analyzování všech knih z učící sady a následně již budou všechny operace určování autorství velmi rychlé, praktický okamžité. Pokud bude chtít uživatel do učící sady přidat, či z ní odebrat nějaké texty, bude mu pro tento účel k dispozici tlačítko, které analyzuje pouze nově přibyté texty a aktualizuje patřičné pomocné soubory.

Program bude schopen operovat ve dvou módech podle druhu trénovací sady. Bude zohledněna situace, kdy máme k dispozici pouze pomocné soubory, získané z předešlé analýzy nějakých textů, avšak nemáme již k dispozici texty samotné. Toto umožňuje sdílet korpus s jinými lidmi a při tom se vyhnout porušování zákona v případě autorským zákonem chráněných knih.

5.2 Získání a porovnávání příznaků

Mezi zkoumané příznaky byla zařazena pouze část těch, které byly zmíněny v kapitole 3. Příčinou omezení množství typů příznaků nebyly technické důvody či nedůvěra vůči vynechaným příznakům, avšak pouze nedostatek času. Mezi vybrané spadají:

- funkční slova
- průměrná délka vět
- velikost slovní zásoby
- *ought to* x *should*
- interpunkční znaménka
- dvojice znaků
- délka slov
- poměry slovních druhů
- dvojice slovních druhů
- omezené trojice slovních druhů

Porovnání bude probíhat odlišným způsobem pro různé charakteristiky. Pro každou bude vyzkoušena řada různých postupů za účelem nalezení pro ni nejúspěšnějšího algoritmu. Plánováno je použití různých variací algoritmu k-nejbližších sousedů v kombinaci s hodnotami mediánu či průměru napříč texty jednotlivých autorů. Zohledněno bude též to, do jaké míry je daný příznak stabilní pro určitého autora. Pokud je směrodatná odchylka hodnot určitého příznaku u nějakého autora nízká a hodnota naměřená pro anonymní text výrazně vybočuje od známých údajů, tak je větší důvod domnívat se, že tato osoba není autorem anonymního textu, než v případě, kdy je směrodatná odchylka vysoká. Toto lze samozřejmě uplatňovat pouze u autorů, pro které máme dostatečné množství trénovacích dat.

Za každou charakteristiku bude autorovi přidělen určitý počet bodů. Osoba s největší sumou bodů bude označena za nejpravděpodobnějšího autora.

Jako počet funkčních slov, jejichž výskyt bude měřen, bylo rozhodnuto začít s počátečním množstvím 200 a toto číslo následně upravovat za účelem nalezení jejich optimální kombinace. Těm, které se v textu vyskytují málo, avšak nejsou příliš běžné a měly by tedy mít velkou výpovědní hodnotu o autorově stylu, bude přiřazena větší váha, než ostatním. Mezi takováto slova by mohlo spadat například „thus“, „whilst“ a řada dalších. Tyto „váhy“ budou realizovány započítáváním každého výskytu takovéhoto slova vícenásobně. Zastoupení jednotlivých funkčních slov bude převedeno do podoby:

$$\frac{hodnota}{počet\ všech\ funkčních\ slov}$$

Délka slov byla zařazena z toho důvodu, že většina prací, zavrhujících tuto charakteristiku, operovala nad výrazně kratšími texty a je možné, že v případě knih bude alespoň ve velmi omezené míře použitelná. Měřeny jsou počty slov o délce 1 - 10 znaků. Každá z těchto 10 získaných hodnot je pak vydělena počtem všech slov v daném textu.

Slovní zásoba bude dána počtem různých slov, nacházejících se na nahodile vybraném úseku o délce 6000 slov. Slova nacházející se na tomto úseku je nezbytné normalizovat, aby slova v různých tvarech byla považována za totožná.

Určování slovních druhů bude též prováděno na omezeném úseku textu, aby se tak zabránilo příliš dlouhé době, kterou by zabralo určení slovních druhů všech slov v celé knize. Pro velké množství knih by jinak doba jejich analyzování stoupla neakceptovatelně vysoko. Délka úseku je stanovena opět na 6000 slov. Všechny algoritmy, pracující se slovními druhy, tak budou operovat pouze na tomto výřezu.

Omezenými trojicemi slovních druhů je míněno měření četnosti výskytu jednotlivých slovních druhů, nacházejících se bezprostředně nalevo a napravo pouze od vybraných slovních druhů. Hledáme tedy pouze takové uspořádané trojice $(a_{1,}a_{2,}a_3)$, kde a_n je slovní druh, pro které platí, že a_2 je specifický námi vybraný slovní druh. Konkrétní slovní druh či několik slovních druhů, které je nejvýhodnější použít jako a_2, budou nalezeny experimentálně.

Mezi sledovaná interpunkční znaménka bylo zařazeno těchto sedm (uvnitř složených závorek): { *! - . ? (: ;* }. Relativní četnost výskytu každého z nich je počítána vydělením množství výskytů daného symbolu celkovým počtem slov.

Dvojice znaků jsou měřeny relativně vůči počtu všech slov v textu. Velká a malá písmena nejsou rozlišena. Kromě písmen jsou brány i číslice, mezery, a jiné znaky.

5.3 Externí nástroje

Některé charakteristiky vyžadují identifikaci slovních druhů. Pro tyto účely byl zvolen open-source program MorphoDiTa[5]. Jeho výhodou je jednoduchá přenositelnost mezi Windows a Linuxem. Používá značky typické pro známý jazykový korpus Penn Treebank[6]. Spouštěn bude automaticky ve fázi extrakce příznaků.

[5] ufal.mff.cuni.cz/morphodita
[6] www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

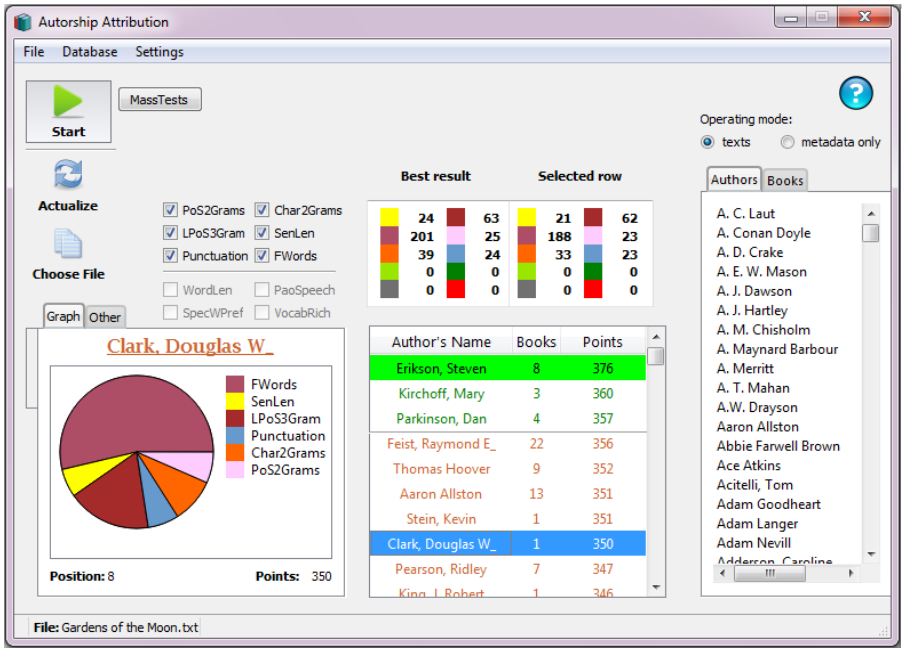
Kapitola 6

Grafické rozhraní

Rozhodnutí pro implementaci grafického rozhraní bylo učiněno ze dvou důvodů. Prvním byla snaha zpřístupnit tento program širšímu okruhu lidí - jediný podobný program, obsahující grafické rozhraní, který je snadno dostupný, má toto rozhraní velmi neintuitivní a nekvalitní. Druhým důvodem bylo to, že grafické rozhraní umožňuje pohodlně vizualizovat řadu informací (například ve formě grafů), díky čemuž je možné si povšimnout vzorů, které by v textovém výstupu šlo snadno přehlédnout. Takto získané informace pak mohou vést k novým nápadům, jak dále zdokonalit jednotlivé algoritmy a dosáhnout lepších výsledků. Naprogramováno bylo v jazyce C++ na frameworku QT.

6.1 Hlavní okno

Použití programu bude probíhat tak, že uživatel nejprve navolí cestu k trénovací sadě. Následně zadá (pomocí dialogového okna) cestu k anonymnímu textu a stisknutím na tlačítko *Start* dojde k nalezení nejpravděpodobnějšího autora a zobrazení výsledků v tabulce. Pokud jsou přidány nebo odebrány nějaké soubory z trénovací sady, stisknutím tlačítka *Actualize* dojde k analýze nových souborů a promítnutí změn do programu.



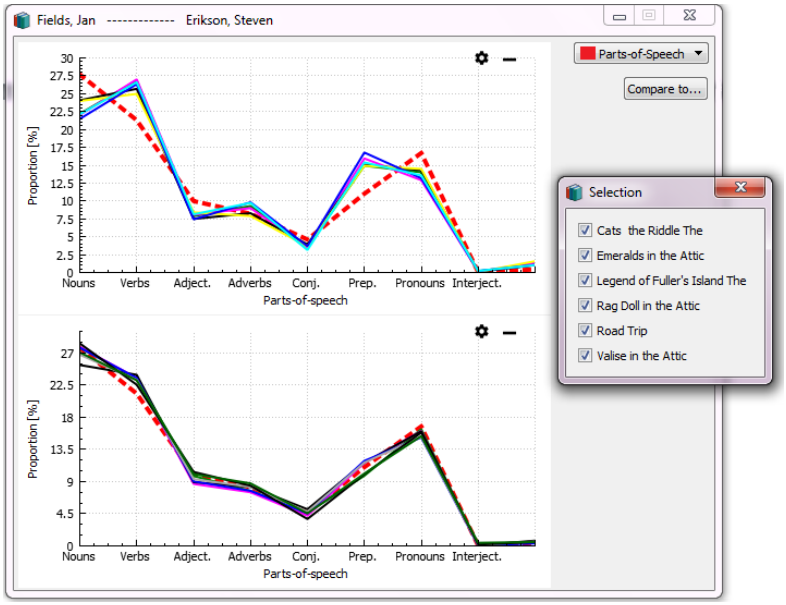
Obrázek 8: Hlavní okno po kliknutí na tlačítko Start.

Koláčový graf zobrazuje podíl jednotlivých charakteristik na výsledném počtu bodů. Pomocí zaškrtávacích kolonek nad ním lze jednotlivé příznaky zapínat/vypínat. Při namíření myši na výřez grafu je pod kurzorem zobrazen počet bodů, přidělených za danou charakteristiku. Tlačítko *MassTests*

spouští plošné testování úspěšnosti určování autorů (více na toto téma v kapitole 7). Většina časově náročných operací je doplněna o ukazatel postupu. Na spodní liště okna lze vidět název aktuálně vybraného textu, jehož autor má být určen. Skrze kliknutí pravým tlačítkem na autora/knihu v seznamu v pravé části okna lze snadno přejít do složky daného autora/knihy.

6.2 Vizualizace charakteristik pomocí grafů

Dvojklikem na výřez grafu v hlavním okně dojde k zobrazení okna s vizualizací hodnot dané charakteristiky. Pro tento účel byla využita knihovna QCustomPlot[7]. Pro různé charakteristiky byly použity různé typy grafů.



Obrázek 9: Okno s vizualizací poměrů slovních druhů

Na obrázku č. 9 lze vidět graf, zobrazující rozložení slovních druhů. Červená přerušovaná čára reprezentuje hodnoty naměřené pro určovaný text, plné čáry reprezentují texty vybraného autora. Kliknutím na patřičné tlačítko lze zobrazit legendu asociující barvy křivek s konkrétními texty. Umožněno je omezit počet zobrazených knih na několik vybraných. Kliknutím na tlačítko *„Compare to“* lze nechat vykreslit druhý graf (pro jiného autora). V případě tohoto obrázku byl takto vespod vykreslen graf pravého autora vybraného textu.

[7] http://www.qcustomplot.com/

Kapitola 7

Testovací model

Pro účely testování byly použity převážně texty stažené z online knižní databáze Project Guttenberg[8]. Doplněny byly o několik textů (resp. příznaků z nich získaných) z půjčovny elektronických knih Open Library[9] a několik knih ze soukromé sbírky. Celkový počet takto nashromážděných textů činí 6120 od 1916 autorů.

7.1 Shromáždění a příprava testovacích dat

Pro účely stažení knih projektu Guttenberg a jejich následnému připravení k použití byla napsána skupina skriptů v jazyce Bash. Knihy byly nejprve staženy pomocí programu *wget*. Následně byly odstraněny všechny duplicitní soubory, lišící se pouze typem kódování. Všechny soubory pak byly rozbaleny a původní archivy smazány. V hlavičkách jednotlivých knih byly dohledány informace o jejich názvech a soubory byly přejmenovány tak, aby tyto obsahovaly místo svého identifikačního čísla. V hlavičkách textů šlo též dohledat jméno autora. Pro každé unikátní jméno autora byla tedy vytvořena složka s tímto jménem a všechny texty pak byly roztříděny do těchto složek. Jména o délce větší než 32 znaků byla zkrácena právě na tuto délku. Knihy napsané více autory, či ty, pro které je autor neznámý, byly smazány. U všech textů bylo smazáno 170 řádků na začátku a 370 řádků na konci, aby došlo k odstranění hlavičky a různých přehledů jmen, kapitol či reklam na jiné knihy. Nakonec byly smazány texty, které byly příliš krátké (obvykle sbírky básní apod.) a manuálně odstraněny ty, u nichž je jako autor uváděn název nějaké instituce (např. „*US Government*“ apod.).

7.2 Způsob testování

Kliknutím na odpovídající tlačítko lze zahájit plošné testování. Jedná se o krajní případ křížové validace, kde trénování probíhá na všech textech, s výjimkou jednoho.V průběhu testování je postupně procházena celá učící sada a pro každého autora, ke kterému jsou nám známy alespoň 2 texty, se provede následující:

1. V paměti se ze seznamu textů daného autora vyjme první dosud nezpracovaný text a umístí se do pomocné struktury. Údaje naměřené pro tento text nejsou nadále nijak asociovány s tímto autorem.
2. U vyjmutého textu je určován autor.
3. Návrat na krok č. 1.

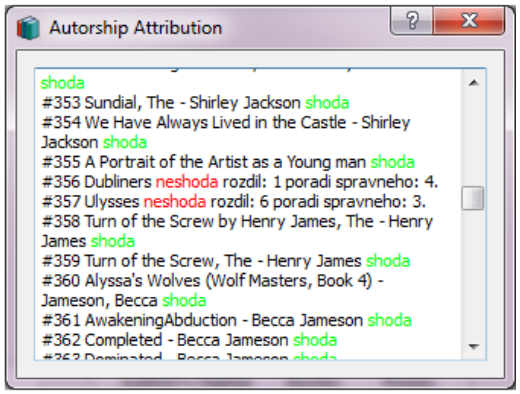
Po dokončení tohoto procesu je zobrazen celkový počet správně a nesprávně určených autorů, z toho odvozený údaj o procentuální úspěšnosti a termínem “multishoda” označený údaj o počtu případů, kdy

[8] https://www.gutenberg.org
[9] https://openlibrary.org

se správný autor umístil na prvním místě, avšak jiný autor obdržel stejně hodně bodů. Tento údaj není brán jako “úspěch” ani jako “neúspěch” a nezapočítává se do procentuální úspěšnosti. Ze způsobu testování je patrné, že trénovací sada bude vždy větší, než testovací sada, neboť takto nelze testovat autory, pro něž známe pouze jedinou knihu.

Kromě procentuální úspěšnosti je v některých případech užitečné používat další ukazatel míry přesnosti. V případech, kdy je autor určen nesprávně, je spočítán rozdíl mezi bodovým ohodnocením skutečného autora a nesprávně označeného autora. Suma těchto rozdílů pak tvoří údaj, jehož nárůst indikuje zhoršení přesnosti.V situaci, kdy úpravou nějakého algoritmu procentuální úspěšnost naroste o několik setin procenta, avšak zároveň dojde k významnému nárůstu této sumy, tak lze většinou konstatovat, že se nejedná o žádoucí úpravu. Tento údaj bude dále v této práci označován symbolem: “$\mathcal{Z}$”.

Významným údajem pro testování je též *maximální chyba*, která vyjadřuje, jak velký byl maximální rozdíl mezi osobou, určenou jako autor textu, a jeho skutečným autorem. V situaci, kdy by byl program v praxi použit k potvrzení autorství něčího textu, lze tuto hodnotu využít k ověření důvěryhodnosti výsledku. Přestože program nemusí určit správně autora ve všech případech, je možné s velmi vysokou pravděpodobností tvrdit, že autorem není nikdo, kdo se umístil bodově ještě dále, než je hodnota maximální chyby, naměřené na dostatečně velkém vzorku dat.



Obrázek 10: Okno zobrazující průběh testování úspěšnosti programu

Díky přehlednému vyobrazení průběhu testování (obrázek č. 10) lze často vidět nárůst či pokles úspěšnosti programu, aniž by bylo nutné nechat jej doběhnout do konce. Toto značně usnadňuje a urychluje experimentování s různými algoritmy pro porovnávání příznaků.

7.3 Rychlost zpracování knih

Na referenčním stroji: Intel® Core™ i5-430M Processor 2.26 GHz, 4GB RAM, připadá na extrakci příznaků z 1000 knih 17 minut a 55 vteřin. Určení autora 1000 knih při výběru z 500 autorů zabere přibližně 19 minut. Určení autorství jediné knihy je velmi rychlé, z hlediska uživatele téměř okamžité. Při maximálním počtu zkoušených autorů trvalo přibližně 2 vteřiny.

Kapitola 8

Vyhodnocení výsledků

Jednotlivé charakteristiky budou v této kapitole uváděny s použitím níže uvedených zkratek. Symbol "*ʒ*" ukazuje relativní velikost chyb a je spolu s termínem *maximální chyba* vysvětlen v předešlé kapitole.

- V - Délka vět
- W - Délka slov
- F - Funkční slova
- Z - Velikost slovní zásoby
- H - Dvojice znaků
- P - Poměry slovních druhů
- D - Dvojice slovních druhů
- T - Trojice slovních druhů s omezením
- I - Interpunkční znaménka
- O - *should* x *ough to*

8.1 Dosažená přesnost

Celá tato podkapitola je strukturována způsobem demonstrujícím postupné přidávání dalších charakteristik v pořadí jejich implementace a jejich jednotlivý přínos pro přesnost programu. Některé charakteristiky jsou natolik nepřesné, že nemá smysl používat je osamoceně, proto zde budou uvedeny pouze v kombinaci s jinými.

8.1.1 Funkční slova

Tento příznak z hlediska úspěšnosti dosáhl jednoznačně nejlepších výsledků. Hodnoty naměřené pro knihy jednoho autora mívají mezi sebou relativně malé odchylky a zároveň umožňují odlišit velké množství autorů. Z tohoto důvodu má tento příznak největší vliv na výsledné rozhodnutí o tom, kdo je autorem. Vyskytly se však případy, kdy 2 knihy od téhož autora měly poměry funkčních slov výrazně odlišné. V tu chvíli je tento přístup nevýhodný, neboť přespříliš špatné hodnocení z tohoto příznaku může převážit nad ostatními charakteristikami.

Tabulka 1: Výsledky při použití příznaku funkčních slov

Použité příznaky	Počet textů v trénovací sadě	Počet autorů	Správně určených	Maximální chyba	*ʒ*
F	2761	1003	80.779 %	105 b	723.5

Výsledků ve výše uvedené tabulce bylo dosaženo s použitím 147 funkčních slov, které byly vyhodnoceny jako nejspolehlivější. Prvotní "draft" 200 slov dosahoval přibližně o 8 % horších výsledků. Jako příklad nejméně spolehlivých slov bych uvedl slova: *was, were, is, are*. Těmito slovy se skrze příznak funkčních slov přenášela informace o tom, v jakém čase hovoří, což se ukázalo být velmi nespolehlivé. Více než dvěma procenty se do výsledku pozitivně promítlo zavedení zvýšené váhy některých méně četných slov.

Nejúspěšnějším postupem pro výpočet bodů, přidělených autorovi za tuto charakteristiku, se ukázalo být následující:

1) Vypočítat průměrnou hodnotu relativní četnosti každého funkčního slova.
2) Vypočítat rozdíl R_1mezi těmito průměrnými hodnotami a četnostmi v anonymním textu:

$$R_1 = \sum_i^{147} |a_i - b_i|$$

3) Stejný vzorec aplikovat na původní (nezprůměrované) hodnoty a vypočítat tak míru odlišnosti jednotlivých knih od anonymního textu. Nejmenší z těchto rozdílů nazveme R_2 .
4) Provést následující výpočet. *C* představuje konstantu, zajišťující, aby se výsledek pohyboval v kladných hodnotách. Konstanty A_1 a A_2 umožňují upravit poměr vlivu hodnot R_k na výsledek.

$$počet\ bodů = C - \frac{R_1}{A_1} + \frac{R_2}{A_2}$$

Vyzkoušeno bylo též použití pouze průměru či pouze nejmenšího rozdílu, nahrazení průměru mediánem či využití směrodatné odchylky, avšak žádný z postupů nevedl k lepším výsledkům, než právě tento.

8.1.2 Délka vět

Až na výjimky byla délka vět napříč autory překvapivě stabilní. Průměrná směrodatná odchylka zde činila 0.42 slova. Přesto však lze nalézt velké množství autorů, u kterých se na tento příznak spolehnout nelze. Nejlépe si zde vedla kombinace mediánu (rozdílů všech knih autora vůči anonymnímu textu) a nejmenšího rozdílu. Jen ve velmi vzácných případech lze však nalézt dva texty téhož autora, jejichž průměrná délka vět by se lišila o více než 5 slov. Pozitivní vliv na úspěšnost tedy mělo přidělení dodatečného počtu bodů všem autorům, jejichž nejmenší rozdíl je nižší, než 5. Úspěšnost výsledného algoritmu v kombinaci s funkčními slovy lze vidět v následující tabulce.

Tabulka 2: Výsledky při použití příznaku funkčních slov a délky vět

Použité příznaky	Počet textů v trénovací sadě	Počet autorů	Správně určených	Maximální chyba	$\mathcal{Z}$
F,V	2761	1003	81.585 %	122 b	771.9

8.1.3 Délka slov

Délka slov se ukázala být zcela nepoužitelná. Všechny pokusy o zařazení tohoto příznaku vedly k poklesu úspěšnosti. Překvapivě špatných výsledků dosáhl i postup, při kterém byla fixní suma bodů přidělena všem autorům, s výjimkou těch, jejichž rozložení délek slov se oproti anonymnímu textu lišilo nejextrémněji.

8.1.4 Velikost slovní zásoby

Příznak počtu odlišných slov, nacházejících se na úseku o fixní délce N slov, se zpočátku zdál být přínosem, neboť velmi mírně navyšoval úspěšnost. Experimenty s různými hodnotami N bylo rozhodnuto o délce N = 6000 slov. Při navýšení počtu autorů až na 1003 však začalo používání tohoto příznaku vést ke zhoršování úspěšnosti a bylo tedy zřejmé, že použití tohoto příznaku je nežádoucí.

8.1.5 Poměry slovních druhů

Zařazení tohoto příznaku mělo jen velmi malý vliv na celkovou úspěšnost, nicméně u takovéhoto množství textů je to nezanedbatelné. Nejlepších výsledků se zde podařilo dosáhnout přidělením bodů dle velikosti nejmenšího nalezeného celkového rozdílu mezi autorovou knihou a anonymním textem. Použití celkového nejmenšího rozdílu bylo úspěšnější, než počítání s nejmenšími rozdíly jednotlivých dílčích slovních druhů, jelikož se tak zachoval vztah mezi četnostmi jednotlivých slovních druhů. Jakékoli zapojení mediánu či průměru do výpočtu, vedlo ke zhoršení výsledků.

Tabulka 3: Výsledky při použití příznaku funkčních slov, délky vět a poměrů slovních druhů

Použité příznaky	Počet textů v trénovací sadě	Počet autorů	Správně určených	Maximální chyba	$\mathcal{Z}$
F,V,P	2761	1003	81.630 %	122 b	775.4

8.1.6 Trojice slovních druhů s omezením

Nejlepších výsledků u tohoto příznaku bylo dosaženo při použití sloves a podstatných slov jako hodnot a_2 (viz kapitola 5). Výpočet byl prováděn zvlášť pro oba typy těchto slovních druhů, jejichž okolí bylo sledováno, a počet bodů pak byl sečten. Použití jiných kombinací slovních druhů vedlo k nižší úspěšnosti. Oba dílčí algoritmy využívají pro výpočet kombinaci mediánu a nejmenšího nalezeného celkového rozdílu (této charakteristiky; mezi autorovou knihou a anonymním textem).

Tabulka 4: Výsledky při přidání příznaku trojic slovních druhů

Použité příznaky	Počet textů v trénovací sadě	Počet autorů	Správně určených	Maximální chyba	$\mathcal{Z}$
F,V,P,T	2761	1003	82.234 %	131 b	782.7

8.1.7 Dvojice slovních druhů

Jedná se již o třetí zařazený příznak, využívající slovních druhů. Jeho přidáním úspěšnost dále stoupla. Pro výpočet je zde používán pouze nejmenší celkový rozdíl. Aplikace mediánu či průměru vedla ke zhoršení výsledků. V tuto chvíli však bylo nutné zjistit, zda-li je skutečně nutné používat všechny 3 tyto příznaky. Byly provedeny experimenty s použitím různých dvojic těchto příznaků a v tabulce níže lze vidět dvě nejúspěšnější varianty.

Tabulka 5: Výsledky při přidání dvojic slovních druhů (*s* a *bez* příznaku poměrů slovních druhů)

Použité příznaky	Počet textů v trénovací sadě	Počet autorů	Správně určených	Maximální chyba	$\mathcal{Z}$
F,V,P,T,D	2761	1003	82.328 %	131 b	780.8
F,V,T,D	2761	1003	82.328 %	131 b	776.7

Z měření vyplynulo, že po přidání příznaku dvojic slovních druhů již nadále není žádoucí sledovat též poměry slovních druhů. Ačkoli procentuální úspěšnost je v obou případech stejná, hodnota celkové chyby $\mathcal{Z}$ je ve druhém případě nižší.

8.1.8 Interpunkční znaménka

Interpunkční znaménka se ukázala být až překvapivě přesným ukazatelem autorova stylu. S nárůstem úspěšnosti o 1.271 % dosáhly většího zlepšení úspěšnosti, než kterýkoli z předešlých příznaků s výjimkou funkčních slov. Z počáteční množiny 7 příznaků (uvnitř složených závorek): { *! - . ? (: ;* } byla odstraněna tečka, neboť razantně snižovala přesnost. Po prohlédnutí řady grafů relativních četností těchto znaků se jevilo jako vhodné odstranit ze sledované množiny též pomlčku, neboť se zde vyskytovaly velké výkyvy v rámci textů téhož autora, nicméně její odebrání vedlo k poklesu úspěšnosti, a tak byla ponechána.

Tabulka 6: Výsledky při přidání příznaku interpunkčních znamének a při jeho použití osamoceně

Použité příznaky	Počet textů v trénovací sadě	Počet autorů	Správně určených	Maximální chyba	$\mathcal{Z}$
I	2761	1003	18.667 %	24 b	266.5
F,V,T,D,I	2761	1003	83.599 %	131 b	760.1

8.1.9 Should vs Ought to

Domněnka, že preferenci mezi modálními slovesy *should* a *ought to* lze spolehlivě použít k identifikaci autora se nepotvrdila. Ačkoli při prvních pokusech na menších počtech autorů jejich použití vedlo k mírně vyššímu počtu správně určených knih, při navýšení počtu autorů na tisíc již tento příznak úspěšnost naopak snižoval.

8.1.10 Dvojice znaků

Dvojice znaků byly jednou z úspěšnějších charakteristik. Výpočet bodů zde probíhal vypočtením rozdílů relativních četností dílčích dvojic mezi anonymním textem a autorovou knihou. Toto se provedlo pro všechny knihy právě zvažovaného autora a na základě velikosti nejmenšího celkového rozdílu pak byl přidělen počet bodů. Nevýhodou tohoto příznaku byl značný nárůst doby výpočtů.

Tabulka 7: Výsledky při přidání příznaku dvojic znaků

Použité příznaky	Počet textů v trénovací sadě	Počet autorů	Správně určených	Maximální chyba	ζ
F,V,T,D,I,H	2761	1003	84.598 %	149 b	784.0

8.2 Škálovatelnost programu

Zásadní vliv na dosaženou přesnost programu mělo množství autorů v trénovací sadě. Se zvyšujícím se počtem autorů razantně narůstá počet případů, kdy více autorů má natolik podobný styl, že je problém je od sebe rozlišit. Aby šlo rozpoznat případy, kdy se správný autor neumístil na prvním místě, avšak nacházel se v žebříčku možných autorů velmi vysoko, oproti případům, kdy se správný autor umístil zcela špatně, bude v následující tabulce uvedena úspěšnost dvakrát - jednou pro případ, kdy je autor identifikován správně a podruhé je jako úspěch započítáván i výskyt mezi pravého autora mezi nejpravděpodobnějšími 10 autory.

Tabulka 8: Porovnání výsledků při výběru z různých počtů autorů

Použité příznaky	Množství textů v trénovací sadě	Počet autorů	Správně určených	Správný autor mezi horními 10
F,V,T,D,I,H	2761	1003	84.682 %	93.964 %
F,V,T,D,I,H	4581	1528	80.519 %	90.633 %
F,V,T,D,I,H	6120	1916	77.577 %	88.532 %

Dodatečnou úpravou, jejímž použitím došlo k nárůstu úspěšnosti (v případě 1003 autorů z 84.59 % na 84.68 %), bylo přidání algoritmu, který pro každý anonymní text hledá 5 knih, které mu jsou nejpodobnější. Pokud alespoň 4 tyto knihy patří stejnému autorovi, tak je mu přidáno nemalé množství bodů navíc.

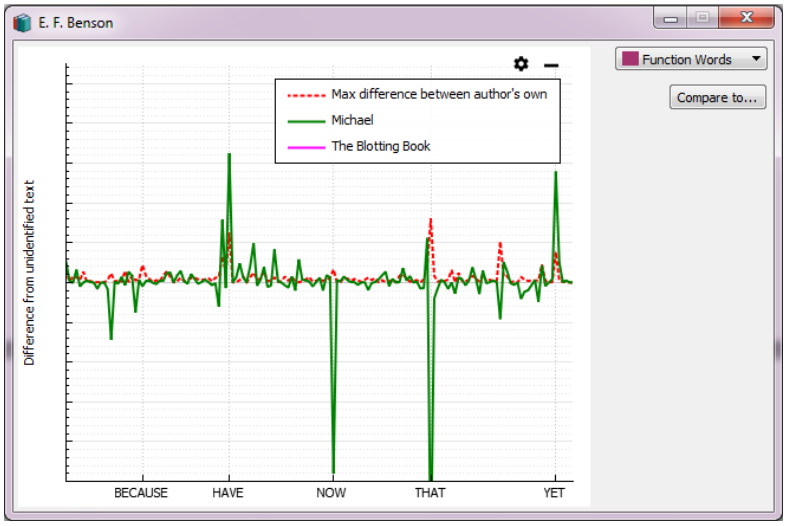
8.3 Analýza chybně určených textů

U některých autorů se nacházejí texty natolik odlišné, že tento program nejenom neurčil správně autorství takovýchto textů, ale dokonce odpovídajícího autora zařadil velice hluboko na setříděném seznamu potencionálních autorů. Tři nejhůře identifikované knihy byly podrobeny bližšímu zkoumání.

Shadows of Flames (Hurst & Blackett, Ltd., London, 1915) od Amélie L. Rives (1863–1945) je první z takovýchto knih. Od této autorky je k dispozici pouze jediná další kniha a při tvorbě autorčina stylu tedy vycházíme pouze z ní. Amélie Rives v tomto případě obdržela extrémně málo bodů za všechny charakteristiky s výjimkou trojic slovních druhů s omezením. Průměrná délka vět obou textů je 11 a 20 slov, což je neobyčejně vysoký rozdíl. Největší propad je však způsoben používáním zcela odlišných funkčních slov. Při bližším pohledu na oba texty je patrné, že zatímco v *Shadows of Flames* se nachází relativně málo archaismů, tak druhá kniha (*A Brother to Dragons and*

Other Old-time Tales (Harper & Brothers, New York, 1888) jich obsahuje velmi mnoho. Tento rozdíl je patrně dán dlouhým odstupem (27 let) mezi napsáním těchto dvou knih. Je pravděpodobné, že pokud by do trénovací sady byly zařazeny některé texty, které publikovala v období mezi těmito dvěma, tak by si program již vedl podstatně lépe.

Druhou vybranou knihou je *Crescent and Iron Cross* (George H. Doran Company, 1918) od E. F. Bensona. Pro tohoto autora bylo trénování provedeno na 2 knihách a všechny tři byly publikovány v rozmezí 10 let. Zatímco obě knihy, na kterých bylo trénování provedeno, měly z hlediska většiny charakteristik velmi podobný styl, tak *Crescent and Iron Cross* po všech stránkách výrazně vybočoval. V grafu funkčních slov (obrázek č. 11) lze vidět dramatický pokles používání slov *that, now, yet* a *have* oproti ostatním jeho knihám. *The Blotting Book* má křívku velice podobnou té zobrazené (zelená), avšak na obrázku je skryta pro větší přehlednost. Průměrná délka věty chybně určeného textu byla 29 slov, zatímco u zbylých dvou knih pouze 15 a 17 slov. Značný rozdíl byl též v používání interpunkčních znamének. Autor zde nepoužil téměř žádné vykřičníky a otazníky. Z bližšího průzkumu samotného textu pak vyplynulo, že zatímco u dvou známých textů se jedná o novely, tak *Crescent and Iron Cross* spadá do literatury faktu. Ačkoli letmý pohled neodhalí žádné dramatické změny oproti jeho ostatním knihám (kupříkladu nějaké dlouhé seznamy údajů apod.), tak přesto se změnou druhu literárního díla přišla i takto výrazná změna autorova stylu.



Obrázek 11: Graf funkčních slov - červená přerušovaná čára představuje maximální odchylku mezi známými texty daného autora. Zelená čára reprezentuje rozdíl mezi knihou *Michael* a neznámým textem. Třetí kniha je skryta.

Třetí z nejhůře identifikovaných knih nese název *Myths of the Norsemen* (George G. Harrap and Co. Ltd., London, 1908). Zde byla příčina odhalena velmi rychle. Jedná se o literaturu faktu a autor zde velmi často citoval jiné osoby. Citací se zde vyskytovalo tak obrovské množství, že znemožnily správně změřit údaje, typické právě pro tohoto autora (množství citací nelze jako určovací rys tohoto autora brát, jelikož druhá jeho kniha neobsahuje žádné).

Kapitola 9

Závěr

Po prostudování technik, používaných pro určování autorství, jsem vytvořil program, který s vysokou pravděpodobností dovede správně identifikovat autora knihy na základě stylometrického rozboru jeho předešlých děl. Tento program dovedl úspěšně přiřadit J. K. Rowlingové knihu *The Cuckoo's Calling*, kterou vydala pod pseudonymem Robert Galbraith. Rovněž dobře si vedl u jiných známých případů (Stephan King, alias Richard Bachman; Isaac Asimov, alias Paul French; a další).

Identifikace autora na počátku probíhala s využitím 10 vybraných charakteristik, testování však odhalilo, že tři z nich úspěšnost zhoršují a posléze byly proto vyřazeny. První ze zmíněných tří je délka slov. Její neúspěch byl očekáván, neboť většina předchozích prací dospěla ke stejnému závěru, nicméně nenalezl jsem žádný případ, kdy by délka slov byla zkoumána u takovéhoto množství takto dlouhých textů, a proto její vyzkoušení mělo svůj význam. Druhou charakteristikou, která se neosvědčila, bylo měření velikosti slovní zásoby pomocí spočítání množství odlišných (normalizovaných) slov na intervalu fixní délky. Poslední z neužitečných charakteristik byla snaha určit, zda autor více preferuje modální slovesu *should*, či *ought to*.

Porovnávání stylů jednotlivých autorů bylo realizováno pro každou charakteristiku zvlášť. Velmi špatných výsledků zde dosahovalo porovnávání hodnot, naměřených pro anonymní text, s průměrem hodnot, naměřených u všech knih právě zvažovaného autora. Naopak velmi dobře si u řady charakteristik vedla kombinace mediánu a hodnot té autorovy knihy, která je anonymnímu textu nejpodobnější. Dodatečného zlepšení úspěšnosti programu se povedlo docílit hledáním 5 knih, které jsou anonymnímu textu nejpodobnější a v případě, že alespoň 4 z nich patří témuž autorovi, tak tento obdrží dodatečný počet bodů (udávajících míru pravděpodobnosti, že právě on je autorem). Překvapivým zjištěním bylo to, že u žádné z charakteristik se mi nepovedlo docílit lepších výsledků zohledněním směrodatné odchylky při výpočtu. Důvodem tohoto však může být to, že pro většinu autorů nebylo k dispozici dostatečně mnoho textů.

Výsledný program obsahuje grafické rozhraní a byl navržen tak, aby jej mohl být schopen použít každý, kdo se o tuto problematiku zajímá. Jednotlivé charakteristiky je možné zapínat/vypínat, snadno lze nechat vykreslit nejrůznější grafy a seznam možných autorů určovaného textu je zobrazen v přehledné tabulce.

Z hlediska dalšího vývoje by bylo vhodné zařadit též několik různých algoritmů strojového učení a pozorovat rozdíly ve výsledné přesnosti programu. Dále je možné zkusit měřit některé další charakteristiky, jako např. frekvence písmen či slov na určitých pozicích.

Literatura

[1] GRIEVE, Jack William: Quantitative Authorship Attribution: A History and an Evaluation of Techniques.
http://summit.sfu.ca/system/files/iritems1/8840/etd1721.pdf, 2002.

[2] JUOLA Patrick: Authorship Attribution.
http://www.mathcs.duq.edu/~juola/papers.d/fnt-aa.pdf.

[3] KARLOVASSI, Samos: Survey of Modern Authorship Attribution Methods.
http://www.icsd.aegean.gr/lecturers/stamatatos/papers/survey.pdf.

[4] RUDMAN Joe, David I. HOLMES, Fiona J. TWEEDIE, R. Harald Baayen SOUVICK a Das DIPANKAR: The State of Authorship Attribution Studies: (1) The History and the Scope; (2) The Problems -- Towards Credibility and Validity.
http://opim.wharton.upenn.edu/~sok/papers/r/s004.html.

[5] BROWN, Steve: Bachman Exposed.
http://www.liljas-library.com/bachman_exposed.php.

[6] ŠABÍK, Matúš.: Určení autorství, bakalářská práce.
https://www.vutbr.cz/www_base/zav_prace_soubor_verejne.php?file_id=119006, 2014.

[7] HAJEK, M.: Neural Networks.
http://www.cs.ukzn.ac.za/notes/NeuralNetworks2005.pdf, 2005.

[8] NG, Andrew: Support Vector Machines.
http://cs229.stanford.edu/notes/cs229-notes3.pdf.

[9] PROMITA Maitra, Ghosh SOUVICK a Das DIPANKAR: Authorship Verification – An Approach based on Random Forest.
http://ceur-ws.org/Vol-1391/134-CR.pdf, 2005.

[10] DIEDERICH Joachim, Jorg KINDERMANN, Edda LEOPOLD a Pass GERHARD: Authorship Attribution with Support Vector Machines.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.7558&rep=rep1&type=pdf.

[11] DOBBS, Sarah: 10 Authors Who Write Under Different Pen Names.
http://mentalfloss.com/uk/books/28108/10-authors-who-write-under-different-pen-names.

[12] STAMATATOS, Efstathios (2009): A Survey of Modern Authorship Attribution Methods. Publikováno v: *Journal of the American Society for Information Science and Technology*, vyd. 60, str. 538-556.